User Manual of *LocalDiff* Version 1.5

Nicolas Duforet-Frebourg and Michael Blum
Université Joseph Fourier,
Centre National de la Recherche Scientifique,
Laboratoire TIMC-IMAG, Grenoble, France.

December 2012

1 Introduction

LocalDiff provides Bayesian measures of Local Genetic Differentiation to characterize non-stationary patterns of isolation by distance. Non-stationary patterns of isolation by distance arise when genetic differentiation between populations (or between individuals) increases at different rates in different regions of the habitat. Typical patterns include barriers to gene flows, secondary contact zone, corridors for gene flow, or gradients of gene flow across the habitat. By analogy with the concept of LocalDiff map in ecology, which measures the cost of movement through the landscape, *LocalDiff* provide estimates of Local Genetic Differentiation with larger Differentiation values indicating larger population genetic differentiation per unit of spatial distance.

2 Algorithm

The software draws a batch of parameter using a MCMC algorithm to characterize decay of a pairwise statistic, and then compute a posterior predictive value of Local Differentiation using equations in [2].

3 Starters

3.1 Download

An archive containing the software can be downloaded at the following webpage:

`http://membres-timc.imag.fr/Michael.Blum/LocalDiff.html`

3.2 Windows OS

If you are using windows (64 bits), you can directly use the software `LocalDiff.exe` . The first one is the command line software. You will have to open a terminal first (`run`, then type `cmd`). Then go in the repertory containing the executable file `LocalDiff.exe`, and just type `LocalDiff.exe` with the parameters of your choice.

3.3 UNIX OS

Extraction and Compilation The archive of the program is provided with a `Makefile` for `UNIX` OS. Compilation proceeds as follows First, you need to compile the local modified Lapack library ([1]).

```
MyMachine $> make lapack
```

Then, compile the program

```
MyMachine $> make
```

After compilation, if for some reasons, you want to clean the repertory of all executables and binary files (including Lapack objects), just type

```
MyMachine $> make realclean
```

If you want to remove all executables and binary files but Lapack objects, just type

```
MyMachine $>make clean
```

After compilation, you can run the program. You can run it without parameters, and a presentation screen will be displayed. Then the software is run as other usual software for LINUX.

3.4 MAC OS

The software has been initially developped for UNIX type of Operating system. It should be running fine with MAC OS.

4 Command line

Here is a complete list of the parameters of the program, and their meaning. When a parameter can be unspecified, it is explicitly mentioned. Basically the command line to run the software is the following one:

```
 MyMachine $> ./LocalDiff -c genetic_Measure MatrixFile datatype
 -i INPUTFile -p PositionFile -o OUTPUTFile
 -l LabelFile -n number_of_neighbors distance_to_neighbors
 -s number_of_posterior_replicates -m doMean -v verbose
-d distance_type
```

-i INPUTFile The input file is the name, with the path, of the file containing the pairwise matrix of correlation, or any other relevant pairwise measure you want to estimate locally.

-c genetic_Measure MatrixFile datatype *LocalDiff* can also handle files of genotypes such as *Structure*'s input files and compute genetic measures of similarity. the first parameter genetic_Measure tells which measure to compute, it can be Cov, Cor for Covariance or Correlation of allele frequencies between populations. Fst for Fst measures as described by Weir and Cockerham between populations. The second parameter tells where to write the calculated matrix. datatype tells what kind of data is in the file, it is either HaploSNP, DiploSNP or MultiAllelic for Microsatellites or others. WARNING: if you use this option, -c is the first parameter you must specify, such as example 3.

-p PositionFile The position file gives the coordinates of all sampled populations or individuals. The positions can either be Cartesian coordinates or the longitude and the latitude of the sampling sites (longitude should be the first coordinate).

-o OUTPUTFile The output file is the name, and path, you want to give to the output file, which contains the locations and LocalDiff values for all samples sites

-l LabelFile The label file may not be specified. It contains the name of the sampled populations or individuals of the data set. Default labels are "pop1, pop2... popn".

-n number_of_neighbors distance_to_neighbors Here are specified the parameters that define the neighborhoods. The first parameter is the number of fictive neighboring populations in the vicinity of each sampled site. The second parameter is the distance between the neighbors and the sampling site. The unit for the distance corresponds to the Euclidean distance when using Cartesian coordinates and is in kilometer when using longitude and latitude. By default, we consider 2 neighbors and we use a distance between neighbors and sampling sites equal to one tenth of the minimum distance between sampling sites.

-s number_of_posterior_replicates Estimates of LocalDiff measures are averaged over posterior replicates of the parameters of the correlogram model. By default, the number of posterior replicates is equal to 100.

-m doMean Set this parameter to 1 if you wish LocalDiff values averaged over unsampled sites and over replicates of the posterior distribution. If however you wish the detail of those values, you can set this parameter to 0. Default value is 1.

-v verbose This parameter specifies the level of details to output from the execution. the parameter is an integer between 0, 1 or 2. Default value is 1.

-d distance_type The coordinates in the PositionFile can be euclidean coordinates, or geographic coordinates, longitudes and latitudes. The distance_type parameter must be chosen accordingly to the coordinate system. The value can be "euclidean", for standard euclidean coordinates, or "great-circle" if positions are geographic coordinates. Default value is "euclidean".

5 Graphical User Interface

If you are not familiar with command line, you may prefer the GUI program of the archive. This Program is a simple friendly user interface that runs *LocalDiff* for you. All the slots to fulfill correspond to arguments of the command line. You can see for example figure 5. To run the software correctly, *LocalDiff* and the GUI must be in the same directory.

6 Files

6.1 Input Files

Similarity Matrix Estimates of LocalDiff measures can be computed for any type of pairwise measures of similarities between populations or individuals, provided that they decrease with geographical distance. Classical measures include the Pearson correlation, one minus F_{st} values, identity by state or by descent measures...

Whatever your choice of statistic the input file should be the same. Each line of the input file corresponds to one row of the matrix, and all features are separated by at least one blank. An example of a 4×4 matrix is provided below.

1	0.79	0.85	0.82
0.79	1	0.80	0.80
0.85	0.8	1	0.89
0.82	0.9	0.89	1

A matrix of larger dimension, which is used in example 1, is provided in `Examples/Matrix1D.dat`.

Genotypes *LocalDiff* can also compute similarity measures from genotypes, and then use those measures in the algorithm. This file must be a $(nsam \times (nbMarkers + 1))$ matrix. The first column corresponds to the population labels, integers from 1 to n. If a LabelFile is used, the labels must be the position of the population label in the label file e.g an individual with label 1 would be of the first population in the labelfile, and so on. The order of the individuals does not matter. Missing values are handled and must be coded with the value -9. Correlation, or Covariance, of allele frequencies between population for SNP data are handled data can be either haploid e.g 0 and 1, or Diploid $0, 1$, and 2. *Fst* and other measures such as Identity-By-State will be available soon.

An example of genotype file is provided below.

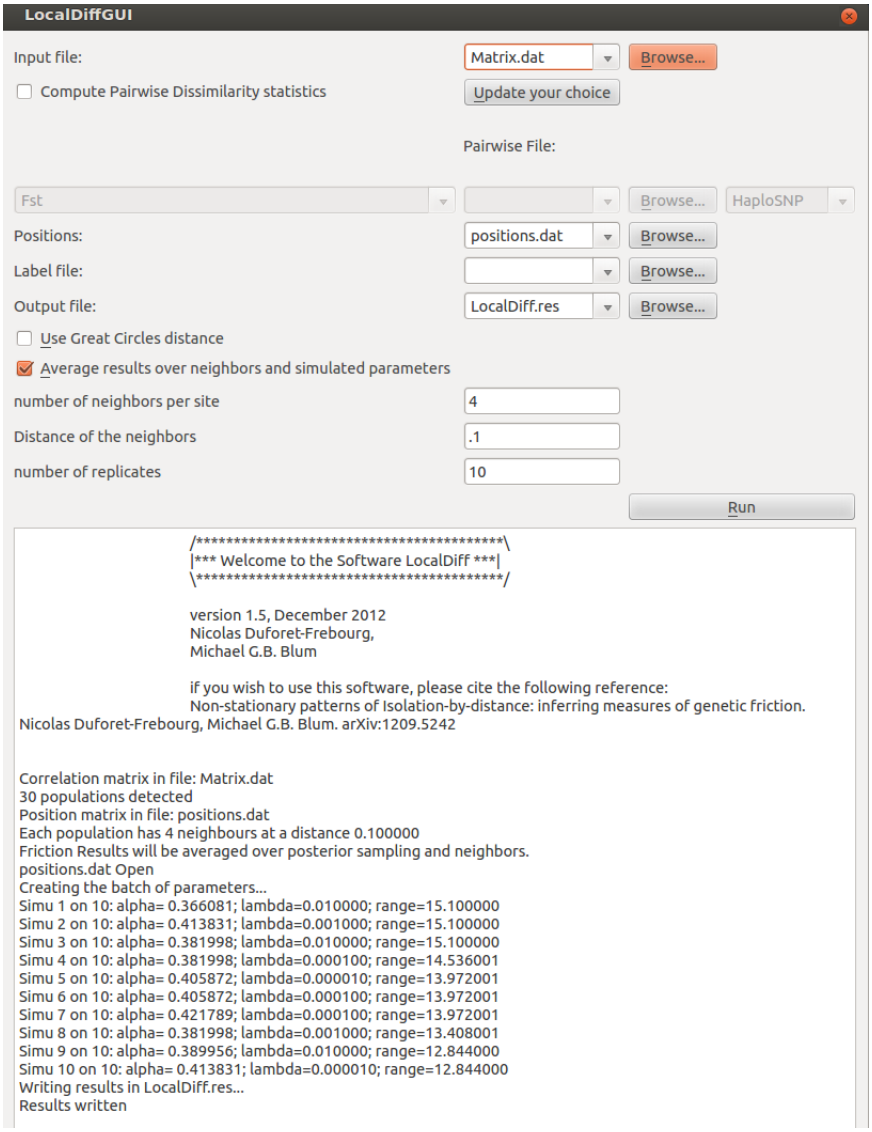

Figure 1: the Graphical User Interface for *LocalDiff*

3	0	0	1	...
1	0	1	−9	...
2	1	−9	0	...
1	1	1	1	...
		...		

A matrix of larger dimension, which is used in example 3, is provided in `Examples/GenoBarrier2D.dat`.

Positions All sampled sites (Populations or individuals) should be geo-referenced. must have an associated geographical Site. The coordinates can either be Cartesian coordinates or geographic coordinates (longitude followed by latitude). Each line of the file corresponds to one sampling site with its associates coordinates. If you are using longitude and latitude, the order matters and longitude must be specified first. Beware, the software checks for number of sites, and number of individuals/populations in the matrix to be the same. If they are different, the program stops with the following message:
`The Number of populations in the position file does not correspond to the dimension of the input matrix`
An example with 4 sampling sites is provided below

1	1
1	2
1	4
1	7

.

The position file of example is provided in `Examples/Position1D.dat`.

Labels The Label file, is an optional input file. It gives names to individuals/populations of your dataset to be printed in the output files. If no name is mentioned, default names would be `pop1, pop2`... To complete the previous example with 4 sampling sites, an appropriate Label file could be:

Michael	*Sean*	*Eric*	*Olivier*

6.2 Output File

For the sack of simplicity there is only one output file after a run of *LocalDiff*. Its name is specified by the `-o` parameter. If one average over replicates and neighbors, this file is a $n \times 4$ array. Every line of the array describes one samples site. The four columns corresponds to the name of the population, the two coordinates, and finally the mean Local Genetic Differentiation.

In the case of a detailed output (-m 0), this output file is an array of dimensions $((n \times n_u) \times (n_{simu} + 4))$ where n is the number of sampled population, n_u is the number of unsampled neighbors by population, and n_{simu} is the number of parameters simulated. Each row corresponds to one unsampled population. The first column is the name of the sampled population whom the unsampled population is the neighbor. The second column is the index of this neighbor for the sampled population. Columns three and four are the coordinates in the habitat of this neighbor. Then the remaining n_{simu} columns are the LocalDiff values, one for each value of parameter simulated. You may want to mean over those values to obtain one Local Genetic Differentiation value per unsampled population, but you can also observe other statistics.

A typical output file with no labels would look like:

*pop*1	1	0.9	1	0.012	...
*pop*1	2	1.1	1	0.013	...
*pop*2	1	1.9	1	0.014	...
		...			

Note: save a logfile Note that if you wish to save a journal of the run, you can still redirect the flow in a log file typing: `MyMachine $> ./LocalDiff ... > myLocalDiffRun.log`

Note: using -c If you are using the pairwise statistic calculation on genotypic data (F_st, Correlations, or Covariances), you will have a second output file. This file is the second argument of the -c option and contain the matrix of pairwise statistics. So you can use *LocalDiff* as a quick way to compute Statistic on your data.

7 Displaying the results with a Local Genetic Differentiation map

7.1 Advocated tools

LocalDiff does not provide any visualization tool for displaying Local Genetic Differentiation map. Thus the software remains really easy to use on any

computer, without calling graphical libraries. Displaying a Local Genetic Differentiation map after a run of *LocalDiff* can be performed with the **R** software, and the packages *sp* and *fields*.

A possibility is to display a map of LocalDiff values using a grid that spans the range of the data. This is done by using another layer of Kriging, in a much more classical way this time. How to display the results with **R** is shown afterwards for two different examples.

8 Examples

8.1 Example 1: 1-dimensionnal habitat with a barrier

In the Example repertory of the archive *LocalDiff.tar.gz* we provide files to run *LocalDiff* on a first simple example. The data were simulated using the software *ms* ([3]). We assume that 30 populations evolved according to a classical stepping-stone model. Five units of coalescent ago, a barrier to gene flow arose between populations 15 and 16. Because of the barrier to gene flow, we expect larger Local Genetic Differentiation measures for populations 15 and 16. The file `Matrix1D.dat` contains the matrix of pairwise correlations of allele frequencies between the 30 populations, and the file `Position1D.dat` contains the coordinates of those 30 populations.

A way to run *LocalDiff* here would be:

```
MyMachine $> ./LocalDiff -i Examples/Matrix1D.dat -p Examples/Position1D.dat
-o Examples/My1DResults -n 2 0.1 -s 200
```

To provide a LocalDiff map, run **R**

```
MyMachine $> R
```

and in the **R** command line, type

```
> source("Rfiles/Display1D.R").
```

8.2 Example 2: 2-dimensionnal habitat with a gradient of migrations

A stepping-stone model was also used for the second example. The habitat is 2-dimensional habitat with a grid (10×10) populations. There is no barrier to gene flow here, but varying effective migration parameter, which decreases from the south-west to the north-east. We expect Local Genetic Differentiation to increase from the south-west to the north-east. The file `Matrix2DG.dat` contains the matrix of pairwise correlations of allele frequencies between the 100 populations, and the file `Position1DG.dat` contains the coordinates of those 100 populations.

The command line for running *LocalDiff* is

```
MyMachine $> ./LocalDiff -i Examples/Matrix2DG.dat -p Examples/Position2DG.dat
 -o Examples/My2DResults -n 4 0.1 -s 200
```

To provide a LocalDiff map, run **R**

```
MyMachine $> R
```

and in the **R** command line, type

```
> source("Rfiles/Display2D.R").
```

8.3 Example 2: 2-dimensionnal habitat with 2 barriers to gene flows

A stepping-stone model was also used for the second example. The habitat is 2-dimensional habitat with a grid (10×10) populations. 2 barriers to gene flow are present here, one between $x = 5$ and $x = 6$ at $T = 5$. The other one between $y = 7$ and $y = 6$, $x > 5$, at $T = 3$. We expect Local Genetic Differentiation to reveal those two barriers. The file `GenoBarrier2D.dat` contains the matrix of genotypes of individuals from 100 populations, and the file `Position1DG.dat` contains the coordinates of those 100 populations.

The command lien for running *LocalDiff* is

```
MyMachine $> ./LocalDiff -c Cor Examples/CorrelationMatrix HaploSNP
-i Examples/GenoBarrier2D.dat -p Examples/Position2DG.dat
-o Examples/My2DResults_2 -n 4 0.1 -s 200
```

To provide a LocalDiff map, run **R**

```
MyMachine $> R
```

and in the **R** command line, type

```
> source("Rfiles/Display2D_2.R").
```

8.4 More detailed plots

Raster file If you want to display the LocalDiff map on a specific region only, you can use a raster file for that. An example of a raster file for displaying the locations above 1,000 meters is given. Generating a LocalDiff map with this raster can be performed by sourcing the file `DisplayFromascFile.R` in **R**.

Administrative Area If your habitat corresponds to an administrative zone, country, county, city... you can use the *global administration areas data base* to restrict the fircion map to the region of interest. How to display the LocalDiff map for the human Swedish sample is shown in `DisplayFromgadmPolygon.R`

References

[1] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen. *LAPACK Users' Guide*. Society for Industrial and Applied Mathematics, Philadelphia, PA, third edition, 1999.

[2] Blum M.G.B Duforet-Frebourg N. Non-stationary patterns of isolation by distance: inferring measures of genetic friction. *ArXiv*, mois 2012.

[3] R.R. Hudson. Generating samples under a wright–fisher neutral model of genetic variation. *Bioinformatics*, 18(2):337–338, 2002.